The Birth of Structural Genomics

J. Berendzen and L. Flaks (P-21); and J. Newman, M. Park, T. Peat, G. Waldo, and T. C. Terwilliger (LS-8)

Introduction

Profound scientific discoveries can change not only the direction and scope of research, but also the way people look at their place in the universe. Consider geography in the time of Magellan, or physics in the early 20th century. Researchers in those times must have viewed the data being produced by new instruments with surprise, wonder, confusion, and elation, for they were the first ones to see the world in a new way. At the end of the 20th century, the streams of data from biology and affiliated disciplines are eliciting a similar mixture of emotions. The discoveries being made in these fields may well mark this page of history as the era in which a comprehensive understanding of the machinery of life finally became possible. This new understanding will become the basis for new technologies and industries and will likely change the way we view ourselves.

Modern molecular biology has developed tools for rapidly determining the complete sequence of DNA bases of an organism, known as its genome (Fig. 1). The revolution in biology is being driven by genomics, and at the moment the exemplary technology of the revolution is the automated DNA sequencer. In 1998, sequencing was completed for genomes of six different microbial organisms, typically pathogenic organisms or ones from exotic environments, each a few megabases in length. By the time you read this, it is likely that the aggregate output of genome sequencing projects worldwide will be greater than one megabase per day and that sequencing a microbial genome will be more routine than a space shuttle flight. In a spectacular accomplishment, the first genomic sequence of an animal—a tiny roundworm called *C. elegans*—was completed in December 1998. Four years from now, it is expected that sequencing of the three gigabase human genome will be complete. With the revolution firmly launched in DNA sequencing, it became clear over the last year to many researchers that it is time to move beyond linear DNA sequences to the three-dimensional (3-D) structures and the functions of the proteins that the DNA encodes.

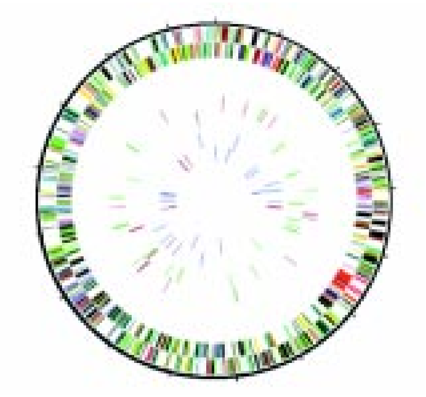

Fig. 1   A microbial genome in map form. Each colored bar represents a different protein.

Protein Structures

Our research centers on the determination and analysis of protein structures. The predominant technology for determining these structures is x-ray crystallography, although multidimensional nuclear magnetic resonance spectrocopy of isotopically-labelled proteins is making an increasing contribution in this area. A structure consists of the 3-D coordinates of the atoms in the molecule (typically several thousand). This information can be abstracted as "cartoons" that show the overall arrangement of the backbone of the peptide chain into elements of secondary structure such as alpha helices or beta sheets (Fig. 2). As beautiful and informative as protein structures are, it is important to be humble about their usefulness. Even in a favorable case where one knows the positions of all the atoms in a protein with high precision (a few hundredths of an Ångstrøm), the structure by itself is only the beginning of an understanding of how the protein works. Yet having a structure, even a low-resolution one, provides a basis for understanding data that would otherwise be difficult to interpret. Using such structures, one can perhaps visualize such phenomena as how substrates dock into the active site of an enzyme, which parts of the protein interact with each other or with other proteins, and which parts of the protein are likely to be floppy.

Proteins are linear polymers of amino acids. There are 20 different kinds of amino acids specified in the genetic code, and the DNA sequence (as determined by genomic projects) specifies the sequence of amino acids that will make up the protein. In turn, the sequence of amino acids determines the 3-D structure of the protein, and the 3-D structure determines the function. Genomic projects are delivering the sequences of a few hundred new proteins every day. One can project that four years from now, we will know the sequences of perhaps 100,000 novel proteins. Determination of the 3-D structures of these proteins is painfully slow by comparison, requiring on average about one man-year of effort by a highly trained scientist at a cost of about $200,000. Tens of thousands of newly-sequenced proteins will nonetheless be attractive targets for pharmaceuticals, agriculture, and chemical manufacturing in the near future. Timely access to their 3-D structures is needed.

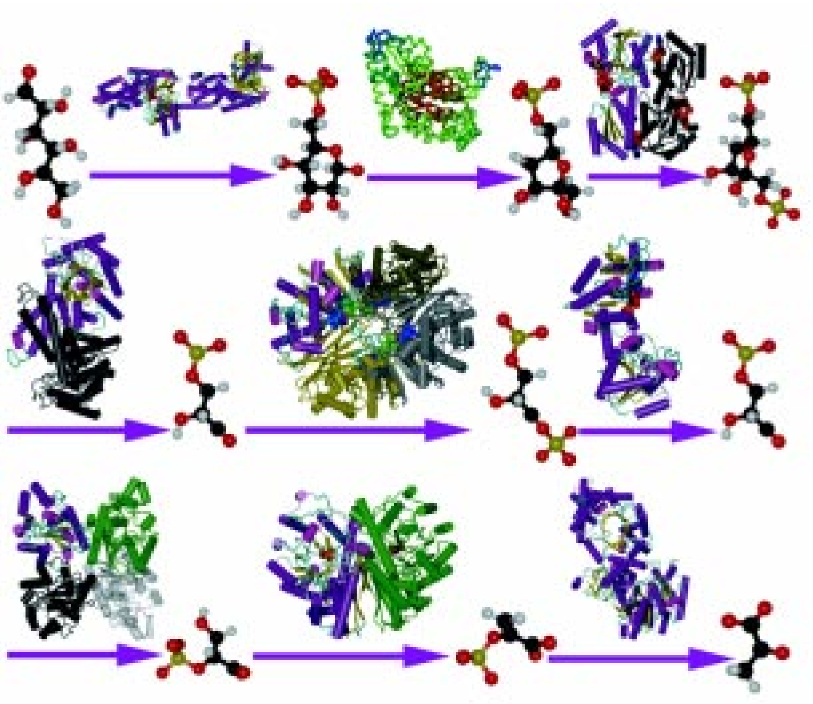

***Fig. 2** Structures enable detailed models of the machinery of life. This figure shows the mechanism of central metabolism, which converts a molecule of glucose into two molecules of pyruvate and stores the energy released as ATP. The cartoons between the molecular models represent the enzymes that carry out the reactions.*

Thus far, researchers have generally appreciated each protein one by one as each structure has been determined, and our understanding of living systems has been crude and incomplete. Currently, however, for well-studied parts of biological systems (such as central metabolism) it is possible to begin building a global and detailed view. If only every pathway in living systems were so well-characterized, how much we might be able to do! After genomic studies pinpointed a protein of interest for medical, chemical, or agricultural applications, then one could use the structural and functional information to move quickly towards a new drug, synthetic process, or disease-resistant plant. It is impossible to conceive of all the uses for such information. Such questions as how genetic variability among patients will affect drug interactions, which at present can only be answered in an approximate way at great expense, might be addressed quickly and accurately through computer simulation if the database that the simulation draws upon is of high enough quality.

Structural Genomics

For the past few years, our team has been working with other structural biologists, genomicists, theorists, and bioinformaticians to determine how best to meet this challenge. A broad consensus has emerged that the structural biology community should mount a large-scale response through an approach called structural genomics. We are encouraged by advances in DNA-sequencing technology developed as part of large-scale genome sequencing projects, and we estimate that a similar large-scale project in structural biology would drive down the time and cost of determining protein structures by an order of magnitude. We believe we should set as our goal the eventual determination of all of the structures of proteins found in nature to an accuracy that is modest at first and improves with time.

The idea behind structural genomics is not simply that the genomes are now available to play with. Nor is it just that there is great potential for improvements in the technology of structure determination and in economies of scale, although those are crucial. Equally important is the idea that through clustering of DNA sequences and 3-D structures, the problem can be divided in a logical way into pieces that research groups can attack and make progress on individually.

Because protein structures fall into families, a structure of one member of the family gives some idea of the structure of the others. How similar the 3-D structures of two proteins will be can be estimated from the similarity of their DNA sequences. At a DNA sequence identity of 25%, for example, the differences in their 3-D sequences can be expected to be about 1 Å rms (root mean square) and the structures will look very similar overall (at the cartoon level). At this level, one could reasonably launch computer modelling efforts to enable drug design, for example. Thus, the structures of the 100,000 proteins thought to be in the human genome could be represented to 1 Å rms by a database of 25,000 structures chosen to give complete coverage at the 25% DNA identity level. If one is willing to start at a much more modest accuracy of 2.5 Å rms (which is accurate enough to assign overall protein fold but not much more), then the number of structures required falls to about 5,000.

To put a scale on the problem, some 500 protein structures have been determined at this level of difference in 40 years of structural biology. Although there has been near-exponential growth in the rate of determination of protein structures, it is clear that to make a serious impact on this problem even at the "fold" level (2.5 Å rms) there will have to be improvements not only in the rate and cost of doing structures, but also in the coordination among research groups in structural biology, bioinformatics, and genomics. At the same time, the traditional focus of structural biology groups on one particular structure determination at a time for a protein of high functional importance will need to be broadened. Any approach that relies heavily on determining one particular protein structure will be prone to getting stuck because that protein may be quite hard to purify or crystallize, while a closely-related structure may be much simpler to determine. Reducing the emphasis on functional importance should permit much faster determination of the structures of one or more members of a broad class of structures.

A Pilot Project

In January 1998, we set out to explore what would be feasible in a structural genomics project given the current level of technology. We wanted to determine the fraction of the proteins in a particular genome that could be rapidly determined using existing methods, identify the bottlenecks in the process, and develop new technology to overcome the bottlenecks. We began by selecting a model organism, *Pyrobaculum aerophilum* (PA), which is a microbe found in undersea vents at temperatures near 100°C. PA has a significant advantage for our studies because it is a hyperthermophile and a member of the ancient *Thermoprotoreales* order of the *Archaea* branch of the tree of life. Because the hyperthermophile proteins are stable at elevated temperatures and those from *E. coli* (a workhorse organism engineered to produce proteins) are not, one step of purifying our proteins could be a simple heat treatment (see Fig. 3). Proteins from hyperthermophiles are also thought to be more stable even at moderate temperatures, and they may crystallize more readily than proteins from a mesophile. PA's membership in the *Thermoprotoreales* order was an advantage because some classes of proteins (such as DNA-processing proteins) from organisms in this branch of the tree of life have a surprising similarity to proteins from humans. PA was also a good choice because its genomic sequence was already being determined by Jeffrey Miller and his team at the University of California at Los Angeles (UCLA), who were eager to collaborate with us on this project.

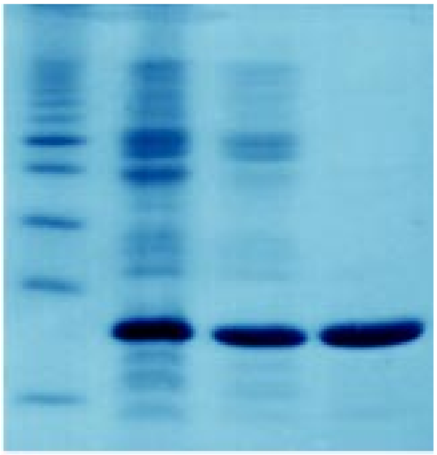

***Fig. 3** Purification of a protein as seen by gel chromatography. The "lanes" of this gel go from low molecular weights (bottom) to high molecular weights (top), with a range of about 40 kDaltons. The lanes from left to right are (1) size standards, (2) crude cell extract from the* E. coli *expression host showing a dark band from high expression of the hyperthermophilic target protein, (3) the same extract after centrifugation, and (4) the same extract after heat treatment at 70°C. Most of the host proteins denature and precipitate after heat treatment, but the hyperthermophilic target protein remains.*

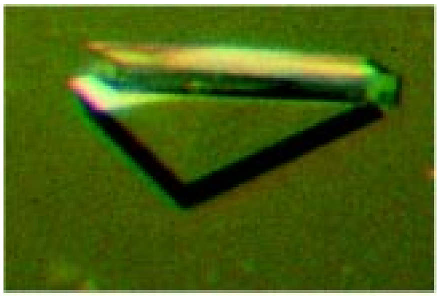

***Fig. 4** A protein crystal from* **Pyrobaculum aerophilum.** *Protein crystals are typically 30 to 70% water, but nonetheless can diffract to atomic resolution. A single protein crystal will often provide enough x-ray diffraction data to solve the structure and determine the positions of each of thousands of atoms in the asymmetric unit of the unit cell.*

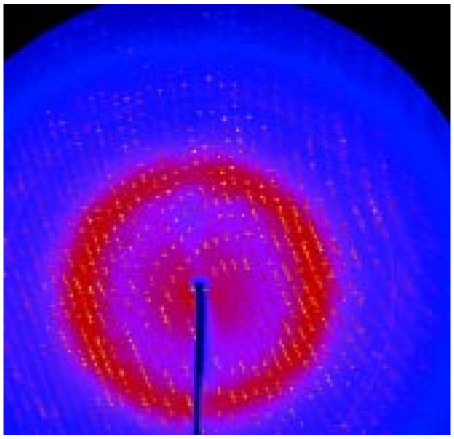

***Fig. 5** In a synchrotron x-ray source, electrons are accelerated to energies of several GeV in a polygonal ring with a circumference of about 1 mile. Radiation at the ring's bends (produced by strong magnets) provides an intense x-ray source that is monochromatized and focused on a protein crystal. The crystal is then rotated over 1° in the x-ray beam to produce a diffraction pattern such as this. The intensities of the spots in this pattern are used in calculating an electron density map. Between 60° and 180° of data like this are typically required for a complete data set.*

Our first step in answering the feasibility question posed above was to eliminate from consideration any proteins that could be anticipated, based on sequence, to be very difficult to solve. This includes proteins that are too big (800 amino acids), proteins that are associated with membranes, or proteins that lack sufficient methionine residues to apply the crystallographic technique of multiwavelength anomalous dispersion (MAD) on selenium-substituted methionines to solve the crystallographic phase problem.

Based on these criteria, we eliminated 60% of the roughly 2,200 genes in the PA genome from consideration. Then, from the remaining genes, we randomly selected a group of 40 proteins and determined how many could be easily expressed (grown) in a production organisim (*E. coli*). Of the 70% of the proteins that could be easily expressed, 40% of these could be easily purified by the conventional techniques of heat-treatment and His-tag affinity chromatography. These purified proteins were subjected to a conventional crystallization screening procedure, in which 30% formed well-diffracting crystals.

SOLVEing the Structure at X8-C

Once we had a protein crystal (Fig. 4), we determined the structure through data collection at a synchrotron facility (Fig. 5). We shipped pre-frozen crystals to the X8-C x-ray crystallography beamline at the National Synchrotron Light Source (NSLS) at Brookhaven National Laboratory. The NSLS is a facility run by the Los Alamos National Laboratory Biophysics Group (P-21) in collaboration with the Canadian National Research Council, Hoffman-LaRoche Pharmaceuticals, the Brookhaven Biology Department, and the Department of Energy (DOE)-UCLA Laboratory of Structural Biology. Typical data collection times at X8-C, which runs 24 hours per day for roughly 210 days per year, are are a few hours per data set.

Using the data collected at the NSLS, we solved the protein structure using MAD phasing and SOLVE, a software package we developed to increase the rate of structure determination (Fig. 6). SOLVE automates the solution of the “phase problem” of crystallography using MAD or multiple isomorphous replacement data. SOLVE is an expert system that uses advanced statistical methods to automatically solve in a few hours a problem that formerly took days or weeks of a trained crystallographer’s time. SOLVE was recognized as one of the 100 most significant inventions of 1998 by *R&D Magazine*, earning the prestigious R&D100 Award (Fig. 8). Once the complete structure of the protein was determined using SOLVE, we refined the atomic positions to produce a final 3-D model of the protein (Fig. 8). Such models allow us to visualize the overall architecture of the protein, and they are the basis for classifying protein structures into families.

Summary and Outlook

The answer to the question posed by our pilot project is that the cumulative percentage of the proteins in the PA genome that could be rapidly determined using existing methods is roughly 10%. At current productivity levels, we estimate that determining the easiest structures would require one to two man-months of effort per structure.

During the past year, we have played a part in the birth of a new field, structural genomics. As with any infant, it is full of potential but most of the development is still ahead. We anticipate that major national and international projects in structural genomics will be launched by the the DOE, the National Institute of Health, and other agencies, and we expect to form a Joint Proteome Institute with our colleagues at the other national laboratories to coordinate our efforts. New technology will have to be developed if the promise of a comprehensive view of protein structures is to be achieved within the next 15 years. Perhaps most importantly, we will have to make models that reach beyond our immediate findings to more and more distantly related proteins. Filling in the gaps in our knowledge will require experimental, theoretical, and computation efforts that are likely to keep the field in a state of excitement for a long time to come.

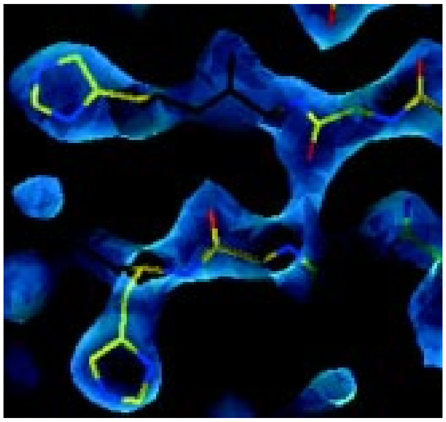

Fig. 6 ***An electron density contour map (blue) and an atomic model of that density (yellow, blue, and red sticks). The process of solving the phase problem in producing the electron density map has been automated by SOLVE.***

Fig. 8 ***The members of the SOLVE R&D 100 award team, Tom Terwilliger and Joel Berendzen.***

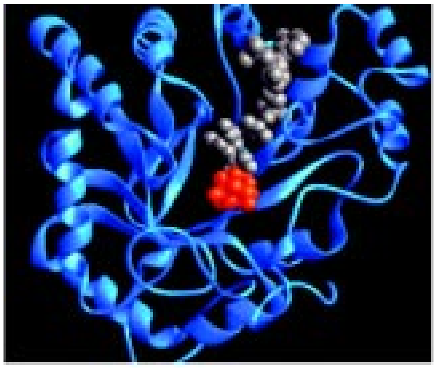

Fig. 9 ***Cartoon of a protein structure showing alternating helical and sheet regions and a space-filling representation of an enzymatic substrate at the active site. Such cartoons illustrate the overall architecture of the protein and are the basis for classifying protein structures into families.***